Neugestaltung der VDE-ITG-Richtlinie zur Bewertung von Kommunikationsendeinrichtungen

Sebastian Möller, Ina Wechsung, Stefan Schaffer, Robert Schleicher, Julia Seebode

Kurzadresse: Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, 10587 Berlin
E-Mail: {sebastian.moeller,ina.wechsung,stefan.schaffer,robert.schleicher, julia.seebode}@telekom.de
Web: http://www.qu-tlabs.tu-berlin.de/

Zusammenfassung

In diesem Beitrag werden laufende Arbeiten zur Neugestaltung der VDE-ITG-Richtlinie 9.4.1.01 zur Bewertung von Kommunikationsendeinrichtungen vorgestellt. Aufgrund des stark veralteten Originals wird als neue Struktur eine Taxonomie von Qualitätsaspekten multimodaler Mensch-Maschine-Interaktion in den Mittelpunkt gerückt. Auf Basis dieser Taxonomie werden verschiedene gängige Evaluierungsmethoden beschrieben, welche neben den „pragmatischen“, funktionalen (*Ease of Use*) auch sog. „hedonische“ Qualitätsaspekte (*Joy of Use*) quantifizieren. Im Anhang werden Beispiele für konkrete Methoden beschrieben. Der Entwurf befindet sich derzeit im Abstimmungsprozess der ITG.

1 Ausgangslage

Die ITG im VDE bietet derzeit eine Empfehlung mit zwei Anhängen zur Bewertung von Kommunikationsendeinrichtungen an: Die Empfehlung 9.4.1.01 „Bewertung von Kommunikationsendeinrichtungen“ beschreibt ein „Prüfverfahren, das dem Entwickler, dem Beschaffer und dem Benutzer hilft, benutzungsfreundliche Kommunikationssysteme zu gestalten bzw. auszuwählen“ [1]. Zusätzlich werden im Anhang [2] von 1997 verschiedene prüfenswerte Funktionen von Telefonen und Anrufbeantwortern sowie im Anhang [3] von 1998 ebensolche Funktionen von Videorekordern aufgelistet. Zu beachten ist, dass sich diese Empfehlung nicht auf die Qualität von über Kommunikationsdienste übertragene Medien (Sprachqualität, Videoqualität) bezieht, sondern auf die Gebrauchstauglichkeit der Endgeräte.

Seit Verabschiedung dieser Empfehlungen haben sich zum einen die zu prüfenden Kommunikationsendgeräte radikal gewandelt. So werden zwar immer noch Telefone und Anrufbeantworter mit ihren Grund- und Zusatzfunktionen angeboten; diese sind jedoch immer häufiger in Bedienabläufe integriert, welche über ein multimodales Portal abgerufen werden können (bspw. eine Bedienoberfläche eines Mobiltelefons, oder ein natürlichsprachlicher Dialog mit einer Netz-Voicebox). Zusätzlich bieten Mobiltelefone als Haupt-Endgeräte von Telekommunikationsdiensten eine Fülle neuartiger Funktionen (SMS, MMS, Web-Browser, Informationsportale, Spiele), welche in der bisherigen Ausgabe der Empfehlung unberücksichtigt bleiben.
Zum anderen sind zunehmend aber auch neue Aspekte der Qualität von Interesse, welche über die funktionalen Anforderungen hinausgehen. Hier sind insbesondere sog. „hedonische“ Aspekte wie die Ästhetik sowie die emotionale Wirkung der Interaktion zu nennen, welche einen wichtigen Aspekt der Gebrauchstauglichkeit (*Usability*) ausmachen und daher auch die Zufriedenheit beeinflussen (vgl. [4]).

2 Neue Grundstruktur

Die Vorbemerkungen machen deutlich, dass die Richtlinie und ihre Anhänge einer gründlichen Überarbeitung und Neugestaltung bedürfen. Um die mit einer Evaluierung erreichbaren Ziele besser zu umreißen wird im neuen Entwurf [5] vorgeschlagen, die Begriffe der Qualität und Gebrauchstauglichkeit zunächst genauer zu beschreiben, als dies in den gängigen Definitionen der ISO geschehen ist. Dies geschieht auf Basis einer Taxonomie von Qualitätsaspekten, welche in Abschnitt 3 beschrieben ist. Anschließend wird eine Klassifikation von betrachteten Anwendungen vorgenommen, zum einen nach dem Funktionsbereich des Gerätes, zum anderen nach den von ihm zur Verfügung gestellten Ein- und Ausgabemodalitäten (vgl. Abschnitt 4).

Der zweite Teil des Empfehlungsentwurfs listet relevante Evaluierungsmethoden auf, welche grob in expertenfokussierte und nutzerfokussierte Verfahren unterteilt werden (Abschnitt 5). Diese werden zunächst bzgl. ihrer Eigenschaften diskutiert und der Anwendungsbereich, Vor- und Nachteile in Form einer Tabelle zusammengefasst. Abschließend wird ein praktischer Leitfaden zur Ableitung geeigneter Testaufgaben sowie zu methodischen Besonderheiten bei der Testdurchführung gegeben. Im Anhang der Neufassung der Empfehlung werden ausgewählte Methoden konkret mit den notwendigen Hilfsmitteln (Fragebögen, etc.) dargestellt, sodass sich der praktische Nutzen der Empfehlung erhöht.

3 Taxonomie von Qualitätsaspekten

Zur Identifizierung relevanter zu prüfender Qualitätsaspekte wird zunächst eine Unterscheidung zwischen der sog. *Quality of Service* (QoS) und der *Quality of Experience* (QoE) vorgenommen, vgl. Abb. 1. Erstere umfasst Indikatoren des Verhaltens und der Leistung von Mensch bzw. System während der Interaktion, sowie die Einflussfaktoren hierauf. Diese lassen sich grob unterteilen in Nutzer-bezogene Faktoren, System-bezogene Faktoren und Kontext-bezogene Faktoren. Die Leistungsindikatoren werden in Form zweier Verarbeitungszyklen dargestellt, einem für den Nutzer und einem für das System.

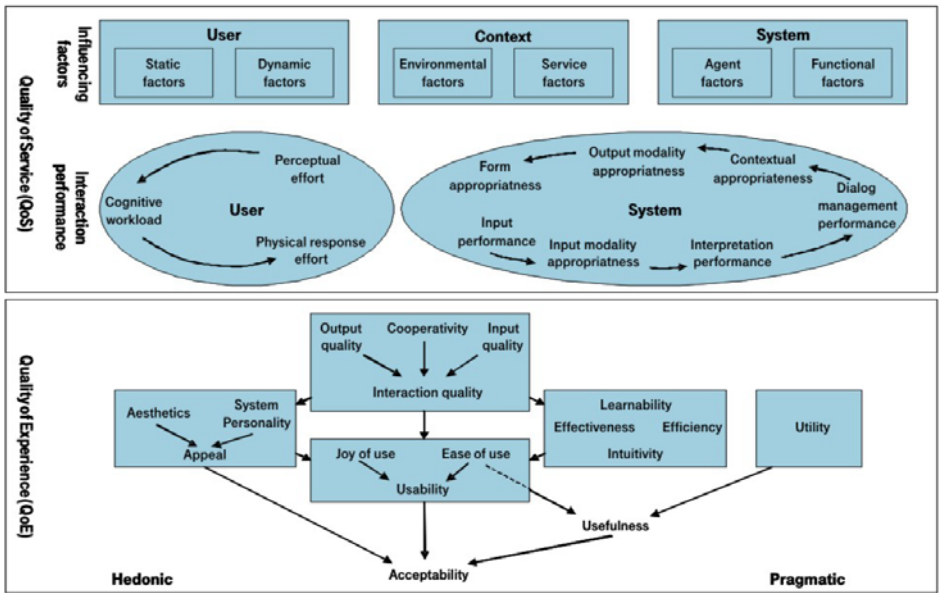

Abbildung 1: Taxonomie der Qualitätsaspekte multimodaler Mensch-Maschine-Interaktion, nach [8].

Letzterem lassen sich eine Vielzahl von Standard-Indikatoren wie die Fehlerrate eines Sprach- oder Gestenerkenners, die Konzeptfehlerrate, die Auftretenshäufigkeit von Korrekturen und Meta-Kommunikation, sowie die Verstehbarkeit von Systemausgaben zuordnen. Auf der Nutzerseite ist vor allem der perzeptive, kognitive und körperliche Aufwand zu nennen, der für die Interaktion notwendig ist.

Im Bereich der *Quality of Experience* wird eine grobe Unterteilung in pragmatische (Funktions-bezogene) und hedonische Qualitätsaspekte vorgenommen. Insbesondere für den ersten Bereich gibt es zwar bereits eine Vielzahl von standardisierten Evaluierungsmethoden, jedoch lassen sich viele nur begrenzt auf die Besonderheiten multimodaler und mobiler Endgeräte anwenden, bzw. sie müssen vor einer Anwendung adaptiert und erweitert werden.

4 Klassifikation von Anwendungen

Diese erfolgt zunächst nach ihrem Funktionsbereich, wobei zwischen Haupt-, Neben- und Zusatzfunktionen unterschieden wird. Daraus werden die folgenden Klassen gebildet: Kommunikationsanwendungen (Telefonie, Videotelefonie, SMS, MMS, etc.), Medienaufzeichnungs- und -abspielanwendungen (Anrufbeantworter, Audio- und Videoabspielanwendungen, etc.) sowie Informationsabrufanwendungen (Informationsportale, Web-Browser, etc.).

Eine weitere Unterteilung erfolgt nach den zur Verfügung gestellten Modalitäten; es wird insbesondere die visuelle, auditive und haptische Interaktion betrachtet.

5 Evaluierungsmethoden

Die Auswahl der beschriebenen Methoden sollte zunächst anhand der Evaluierungsfragestellung sowie anhand der Stelle im Entwicklungszyklus entschieden werden, an der die Evaluierung durchgeführt wird. Folgende Fragestellungen stehen dabei üblicherweise im Fokus [6], [7]:

- Die Evaluation ist auf ein bestimmtes Anwendungsziel ausgerichtet, die Fragestellung lautet: „Ist das System hierfür gut genug?“
- Die Evaluation vergleicht zwei alternative Designs, die Fragestellung lautet: “Welches System ist besser?“
- Mit der Evaluation soll der realen Welt näher gekommen werden: „Wie gut arbeitet das System in der realen Welt?“
- Mit der Evaluation wird die Einhaltung von Standards überprüft: „Wie gut erfüllt das System den/die Standards?“

Bezüglich des Stadiums der Systementwicklung unterscheidet man üblicherweise in:

- Formative Evaluation: Prozessorientierte Evaluation, die schon vor dem Vorhandensein eines Prototyps, also schon im Designstadium, erfolgen kann.
- Summative Evaluation: Ist ergebnisorientiert und erfolgt als eine zusammenfassende Beurteilung des fertigen, schon implementierten Systems.

Auch das Level der erhobenen Informationen ist zu beachten, So können Informationen die durch ein Evaluation gewonnen werden, Informationen von niedrigem Level (z.B.:“Welche Buttons sind am verständlichsten“) oder Informationen von hohem Level (z.B.: „Ist das System leicht zu benutzen?“) umfassen.

Weitere Differenzierungsmerkmale sind der Grad der Kontrolle über das Experiment und der Grad der Nutzerbeteiligung. Für die expertenfokussierte Evaluierung ergeben sich daraus die Verfahren *Cognitive Walkthrough*, Prüflisten, heuristische Evaluierung und modellgestützte Evaluierung, sowie für die nutzerfokussierte Evaluierung die Verfahren Experimente, Interviews/Fragebögen sowie Protokollanalysen, vgl. folgende Abschnitte.

5.1 Cognitive Walkthrough

Der *Cognitive Walkthrough* untersucht die Adäquat-

heit einer Anwendung zur Erledigung einer bestimmten Aufgabe. Dabei analysieren Experten – üblicherweise Usability-Experten, Designer oder Psychologen – die Funktionalität des Systems auf Basis einer Beschreibung der Schnittstelle, einer Beschreibung der Aufgabe, welche der Nutzer ausführen wird, einer Liste von notwendigen Handlungen, um die Aufgabe zu erfüllen, sowie von Informationen über den Nutzer und den Nutzungskontext [9]. Der Experte geht die zur Erledigung der Aufgabe notwendigen Schritte durch und überprüft bei jedem Schritt, ob der Nutzer diesen Schritt wahrscheinlich auch so ausgeführt hätte. Abweichungen werden notiert und zur Optimierung des Systemdesigns eingesetzt.

Vorteil des *Cognitive Walkthrough* ist die Unabhängigkeit von Endnutzern und einem voll funktionstüchtigem System; das Verfahren kann also schon sehr früh im Entwicklungsprozess eingesetzt werden. Der Designer wird unterstützt, die Perspektive eines potentiellen Nutzers einzunehmen; hierdurch können Nutzerziele und Vermutungen der Nutzer über das System definiert werden [10]. Nachteile dieses Verfahrens sind das niedrige Niveau der resultierenden Informationen. Auch kann die Durchführung abhängig vom zu testenden System sehr langwierig sein und bei falscher Aufgabenselektion verzerrte Ergebnisse liefern [10].

5.2 Prüflisten

Mit Hilfe von Prüflisten soll die Evaluation von Kommunikationsendgeräten durch Experten sowohl aufgabenunabhängig als auch unter Betrachtung konkreter Aufgaben ermöglicht werden. In einer gegliederten Liste werden dazu Eigenschaften zusammengestellt, die vom betrachteten System gefordert werden. Viele Prüflisten bieten lediglich eine Aufzählung solcher Eigenschaften in Form von Gestaltungsregeln; eine zusätzliche Bewertung durch Angaben über die Erfüllung einzelner Kriterien und deren Gewichtung durch die Prüfer ist aber möglich. Bei der Gestaltung von Interaktionen muss sich insbesondere die Gestaltung des Dialogs an Aufgaben orientieren. Ergonomische Kriterien, die die menschliche Sensorik und Motorik betreffen, werden dagegen als weitgehend aufgabenunabhängig betrachtet.

In [1] wurde eine Prüfliste erarbeitet, die sich in zwei Hauptteile – einen aufgabenunabhängigen und einen aufgabenorientierten Teil – gliedert. Ähnlich den Prüflisten ist das Verfahren der heuristischen Evaluation (s. Abschnitt 5.3) sehr breit angelegt, jedoch sind Prüflisten in der Regel ausführlicher.

5.3 Heuristische Evaluation

Die heuristische Evaluation ist eine Methode des *Discount Usability Engineering*, ein von Nielsen eingeführter Ansatz, mit pragmatischer und kostengünstiger Ausrichtung [11].

Bei der heuristischen Evaluation bewerten mehrere Experten die Benutzerschnittstelle dahingehend, ob gängigen Usability-Prinzipien (Heuristiken) Rechnung getragen wird [11]. Jeder Experte bewertet die Schnittstelle für sich allein und führt unter Betrachtung der Heuristiken mehrmals Interaktionen mit der Schnittstelle durch. Erst nach Abschluss aller Bewertungen kommunizieren die Experten miteinander, um ein aggregiertes Urteil zu treffen. Durch ein möglichst unabhängiges, unverzerrtes Urteil sollen möglichst viele verschiedene Usability-Probleme aufgedeckt werden. Das Ergebnis einer heuristischen Evaluation ist eine Liste von Usability-Problemen, die genau beschrieben und begründet sein sollen [11]. Weiterhin können die Probleme hinsichtlich ihrer Häufigkeit, ihres Einflusses und ihrer Persistenz beurteilt werden. Gewöhnlich werden ca. 3-5 Experten benötigt, die ca. 60-70 Prozent der Probleme aufdecken. Bei mehr als 10 Experten sind kaum verbesserte Ergebnisse zu erwarten [11].

Die Vorteile der heuristischen Evaluation liegen in der einfachen, zeit- und kostengünstigen Durchführbarkeit sowie in der universellen Anwendbarkeit während des gesamten Designprozesses [10]. So ist auch eine Bewertung von nicht implementierten Systemen anhand des Systementwurfes möglich [11]. Der Nachteil ist, dass die heuristische Evaluation (wie andere Experten-Evaluationen auch) nur angenommene, und nicht unbedingt tatsächlich vorkommende Nutzerprobleme erfasst.

5.4 Modellbasierte Evaluation

Die modellbasierte Evaluation erfolgt auf Basis von Modellen, mit denen Nutzerverhalten während der Interaktion mit dem System beschrieben werden kann. Viele solcher Modelle basieren auf Theorien der kognitiven Psychologie; Beispiele hierfür sind GOMS [12], die *Cognitive Complexity Theory* (CCT) [13], *Adaptive Control of Thought–Rational* (ACT-R) [14], oder *State, Operator And Result* (SOAR) [15]. Andere Modelle basieren auf probabilistischen Zustandsautomaten, die mögliches Nutzerverhalten über Wahrscheinlichkeiten beschreiben. Die Wahrscheinlichkeiten werden entweder über Regeln gesteuert, die Nutzer- und Systemeigenschaften widerspiegeln [16], oder sie werden aus empirisch ermittelten Daten trainiert [17].

Die modellbasierte Evaluation kann schon sehr früh im Designprozess angewandt werden und ermöglicht durch die theoretische Fundiertheit einen tiefen Einblick ins Verhalten der Nutzer [10]. Der Nachteil liegt in dem erforderlichen Zeitaufwand und dem benötigten, sehr hohen Grad an Expertise zur Erstellung der notwendigen Modelle; letzterem kann durch eine Integration des Modells in die zur Erstellung der Anwendung notwendigen Werkzeuge begegnet werden.

5.5 Experiment

Im Gegensatz zur expertenfokussierten Evaluation werden bei der nutzerfokussierten Evaluation tatsächliche – und nicht nur angenommene – Probleme analysiert. Diese Verfahren besitzen also potenziell eine höhere Validität. Beim Experiment werden spezifische Aspekte des Interaktionsverhaltens eines Systems kontrolliert bewertet. Hierdurch ist es möglich, Daten hoher Qualität zu erheben, da Störvariablen im hohen Maße ausgeschaltet oder kontrolliert werden können. Im Gegenzug besteht jedoch die Gefahr, dass die Ergebnisse nicht das Nutzerverhalten in natürlichen Umgebungen abbilden. Auch ist

die Planung, Durchführung und Auswertung eines Experimentes zeit- und kostenintensiv.

5.6 Interviews / Fragebögen

Der Einsatz von Fragebögen und Interviews bietet sich an, wenn Daten erhoben werden sollen, die über Performanzdaten hinausgehen. Insbesondere zu Erfassung „subjektiver“ Daten wie Nutzerzufriedenheit, Einstellungen und Ängste der Nutzer wird auf Fragebögen und Interviews zurückgegriffen. So können Konstrukte erfasst werden, die sich anderen Verfahren entziehen.

Der mit Interviews und Fragebögen verbundene Aufwand sowie die zu erwartenden Ergebnisse hängen stark vom Kontext ab, in dem sie erhoben werden. Beim Einsatz innerhalb eines kontrollierten Experimentes ist der Aufwand durch das Experiment bestimmt; der Fragebogen oder das Interview selbst erfordern nur einen geringen Mehraufwand. Allerdings ist das Ergebnis dann auf die Laborsituation beschränkt. Fragebögen oder Interviews in einem Feldtest können hier validere Ergebnisse bringen.

5.7 Protokollanalysen und Thinking Aloud

Bei Protokollanalysen werden Nutzer und Systemverhalten mittels Video-, Audio-, Log-Dateien oder schriftlichem Protokoll festgehalten. Die wohl bekannteste Methode ist hierbei das *Thinking Aloud*, das laute Denken. Die Testpersonen werden während der Systemnutzung dazu angehalten, ihr Verhalten und ihre Gedanken zu verbalisieren und laut zu äußern [10]. Dies kann während des Tests geschehen oder nachträglich als retrospektives lautes Denken, bei dem der Nutzer (Video-) Aufzeichnungen des Testablaufes kommentiert. Bei letzterem ist jedoch mit Verzerrungen durch Erinnerungsfehler zu rechnen.

Das laute Denken kann sowohl bei freier Exploration der Schnittstelle als auch bei der Bearbeitung konkreter Aufgabenstellungen angewandt werden. Nutzungsprobleme lassen sich damit leicht detektieren. Jedoch wird *Thinking Aloud* manchmal als unnatürlich und verwirrend empfunden wird; manchen Probanden fällt das detaillierte Verbalisieren ihrer Gedanken schwer. Hinzu kommen Probleme bei Schnittstellen, die auditive Ein- und Ausgaben verwenden (z.B. Sprachdialogsysteme), da diese Informationen durch das ununterbrochene Verbalisieren verdeckt werden können.

6 Schlussbemerkung

Der aktuelle Entwurf [5] wird derzeit im FB2 der ITG diskutiert und soll bis Ende 2010 verabschiedet werden. Hierbei soll insbesondere ein detaillierter Anhang mit konkreten Beschreibungen der Evaluationsmethoden erstellt werden.

Literatur

[1] VDE-ITG-Richtlinie 9.4.1.01 (1995). *Bewertung von Kommunikationsendeinrichtungen.* VDE-Verlag, Berlin/Offenbach.

[2] — (1997). *Anhang für Telefone* und *Aufgabenliste Anrufbeantworter*.

[3] — (1998). *Anhang für Videorekorder*.

[4] M. Hassenzahl, M. Burmester und F. Koller (2003). *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität*, in: G. Szwillus, J. Ziegler (Hrsg.): Mensch & Computer 2003, Stuttgart: Teubner, 187-196.

[5] I. Wechsung, S. Schaffer, J. Seebode, R. Schleicher und S. Möller (2010). *Bewertung von Kommunikationsendeinrichtungen*. Neuentwurf der VDE-ITG-Richtlinie 9.4.1.01, TU Berlin.

[6] A. Dix, J. Finlay, G. Abowd und R. Beale (2004). *Human-Computer Interaction.* Hillsdale: Prentice Hall.

[7] J. Preece, Y. Rogers., H. Sharp, D. Benyon, S. Holland, und T. Carey (1994). *Human-Computer Interaction*. Wokingham: Addison Wesley.

[8] S. Möller, K.-P. Engelbrecht, C. Kühnel, I. Wechsung und B. Weiss (2009). *A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction*, in: First International Workshop on Quality of Multimedia Experience (QoMEX'09), July 29-31, San Diego CA.

[9] C. Wharton, J. Rieman, C. Lewis und P. Polson (1994). *The Cognitive Walkthrough: A Practitioner's Guide*, in: J. Nielsen and R.L. Mack (Hrsg.), Usability Inspection Methods, New York: Wiley & Sons.

[10] A. Holzinger (2005). *Usability Engineering for Software Developers*, Communications of the ACM 48(1), 71–74.

[11] J. Nielsen (1994). *Cost-justifying Usability*, Boston: Academic Press, .

[12] S. K. Card , A. Newell und T. P. Moran (1983). *The Psychology of Human-Computer Interaction*, Mahwah: Lawrence Erlbaum.

[13] D. E. Kieras und P. G. Polson (1985). *An Approach to the Formal Analysis of User Complexity*, International Journal of Man-Machine Studies 22, 365-394.

[14] J. R. Anderson und C. Lebiere (1998). *The Atomic Components of Thought*, Mahwah: Lawrence Erlbaum.

[15] A. Newell (1990). *Unified Theories of Cognition*, Cambridge: Harvard University Press.

[16] S. Möller, R. Englert, K. Engelbrecht, V. Hafner, A. Jameson, A. Oulasvirta, A. Raake und N. Reithinger (2006). *MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations*, in: Proc. 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh PA, 1786-1789.

[17] S. Möller, R. Schleicher, D. Butenkov, K.-P. Engelbrecht, F. Gödde, T. Scheffler, R. Roller und N. Reithinger (2009). *Usability Engineering for Spoken Dialogue Systems Via Statistical User Models*, in: 1st Int. Workshop on Spoken Dialogue Systems Technology (IWSDS 2009), 9-11 Dez., Kloster Irsee.